PeopleMap: NLP and Visualization Tool for Mapping Out Researchers

Undergraduate Researcher Name:

Printed: Jon Saad-Falcon

Signature: JSF

Faculty Member #1:

Printed: Duen Horng (Polo) Chau

Signature: Polo Chau

Faculty Member #2:

Printed: Diyi Yang

Signature: Diyi Yang

Table of Contents

Abstract

Discovering research expertise at universities can be a difficult task. Directories routinely become outdated, and few help in visually summarizing researchers' work or supporting the exploration of shared interests among researchers. This results in lost opportunities for both internal and external entities to discover new connections, nurture research collaboration, and explore the diversity of research.

To address this problem, at Georgia Tech, we have been developing PeopleMap, an open-source interactive web-based tool that uses natural language processing (NLP) to create visual maps for researchers based on their research interests and publications. Requiring only the researchers' Google Scholar profiles as input, PeopleMap generates and visualizes embeddings for the researchers, significantly reducing the need for manual curation of publication information. To encourage and facilitate easy adoption and extension of PeopleMap, we have open-sourced it under the permissive MIT license at https://github.com/poloclub/people-map. PeopleMap has received positive feedback and enthusiasm for expanding its adoption across Georgia Tech.

Introduction

University directories provide a tool for faculty, staff, students, and others to find individuals in different colleges, departments, research labs, and fields of study. For the most part, these directories provide a simple classification that places all the individuals at a university in their respective roles with an associated list of characteristics. However, for academic researchers specifically, they often classify a professor as being in a certain department or field of study, when he or she teaches courses in an entirely unrelated subject. Additionally, there are often professors in a specific department (e.g. computer science) that pursue research in a subject more strongly associated with another field (e.g. bioinformatics within the field of biology).

Therefore, this misclassification is often a frustration for companies, corporations, and governmental agencies that seek to pursue business or fund projects at a specific university. It becomes difficult for these organizations to effectively search for a researcher pursuing a desired field when they have to go through a directory that has both false and incomplete information related to all the academic researchers stored. This hinders the ability of both external individuals and internal individuals to accurately locate research labs, understand the topics of study for different researchers, and find individuals related to different fields of study.

Considering this misinformation as well as the desire from various groups of people to better understand this information, we have developed PeopleMap, an interactive tool that "maps out" researchers based on their research interests and publications by leveraging embeddings generated by natural language processing (NLP) techniques. PeopleMap provides the following contributions:

- PeopleMap serves as the first visualization dedicated to helping users explore researcher embeddings; while there has been research that develops methods to recommend research papers and publication venues (Alhoori 2017, Beel 2017, Beel 2016, Küçüktunç 2013, Medvet 2014), less work focuses on developing usable easy-to-access tools for users to interactively explore researcher datasets. PeopleMap fills this research gap and seeks to improve the interpretability and explorability of researcher datasets.
- PeopleMap also provides an open-source, sustainable web application for the community that can be easily accessed via web browsers and implemented as a web-based application. PeopleMap is registered under the permissive MIT license, and its code repository is available at https://github.com/poloclub/people-map. Besides the PeopleMap visualization, it also provides a series of data collection and preprocessing tools that allows users to create a researcher dataset from any list of researchers found on Google Scholar. Additionally, it includes a step-by-step documentation guide (https://app.gitbook.com/@poloclub/s/people-map/) that covers every step of the process from downloading the repository to launching the PeopleMap platform. With the combined data collection resources and PeopleMap visualization, the tool provides an automated solution for researcher interest summarization and discovery, which simplifies the exploration of the work of scientific researchers.

Literature Review

The advent of natural language processing (NLP) has allowed a variety of new industries to develop more interactive and sophisticated tools for analyzing information and conveying connections within data. From finance to engineering to anthropology, the realm of NLP is expanding and finding applications in previously unconsidered places. One key location of this growth is within the field of data visualization, where the goal of conveying important trends within a data set can often be quite difficult.

The difficulty within data visualization in the field of NLP is that the intricacies of language data sets are often more complex and intractable than people think. Before the field of natural language processing emerged, linguistics gave us a human-generated understanding of language through grammar, syntax, and etymology. However, the field of linguistics has experienced difficulties when analyzing large data sets, finding deeper levels of context, and exploring larger connections between words. This divide is even more apparent with the novel advances in the field of NLP with new algorithms, such as BERT (Devlin et al., 2018), that can be configured to cover complex forms of contextual analysis. Furthermore, NLP seeks to account for the inherent bias and failures in language models (Dubossarsky, 2017). Overall, the field of NLP incorporates the new developments in deep learning into more sophisticated and advanced models for language.

However, NLP itself has encountered its own set of challenges in the analysis of language. The field has emerged and grown in popularity since the 1990s when the increased computation power made neural networks, and by extension, deep learning, possible at reasonable speeds. As a result, the 2000s and 2010s have seen a rapid growth in NLP algorithms

but increased computation ability doesn't equal better models for human language. The techniques of bag-of-concepts (Kim, 2017) and non-negative matrix factorization (Lee, 2001) can extract larger topics and ideas from corpora as diverse as healthcare and political science. However, due to their inherent complexity, the models can exaggerate noise in the data and create patterns where they don't exist; this is especially problematic when visualizing NLP data using T-SNE (Maaten, 2008) since it can lead to conclusions about data sets that are not reproducible. Even though we have found more effective techniques for finding data correlations in recent years through the development of word vectors (Xing et al., 2014), there are still several frontiers within the fields of information extraction and summarization.

Considering the gaps within the contrasting views of linguistics and NLP on both data analysis and visualization, the goal of advancing our understanding of language will have to incorporate aspects from both fields while also mitigating the issues found within either view. The power of linguistics lies in its ability to use the centuries of human literary and linguistic analysis to generate vast and thorough rule-based algorithms that can process idioms, translations, and expressions in ways that are often difficult for artificial intelligence (Sanz et al., 2013). To this day, these rule sets are often the best classifiers for the simple nuances of language since they easily map these figurative forms of language to different meanings and contexts. Within the realm of NLP, however, several mathematical techniques have proven especially effective in a huge variety of language datasets. The matrix decomposition called term frequency-inverse document frequency (TFIDF) is ubiquitous in the field of NLP due to its simple yet effective analysis of language (Aizawa, 2003); it is often the only technique within NLP needed for finding the different topics found within some corpora. When combined with

general clustering techniques such as K-means clustering (Likas, 2003), the patterns within NLP datasets can become incredibly clear and easy to investigate.

By combining these two disparate fields, it is clear that the field of data visualization has already begun improving our understanding of language. In the subfield of summarization, NLP and linguistics are being combined to parse documents for key pieces of information that can be concatenated together into bulleted summaries for the sake of thinning document fluff within patient discharge summaries (Krauthammer, 2001); in this data tool, the techniques of bag-of-words from NLP is combined with medical terminology linguistics to create a doubly effective tool. Within the field of journalism, the same trend is occurring; NLP-driven algorithms are being combined with linguistics databases of key political interest languages (such as Russian and Mandarin) to analyze international alliances between the U.S. and its regional parts throughout the world (Tannier, 2016). Rather than becoming opposing fields, NLP and linguistics are combining to become an even more potent combination in analyzing human communication.

Methodology

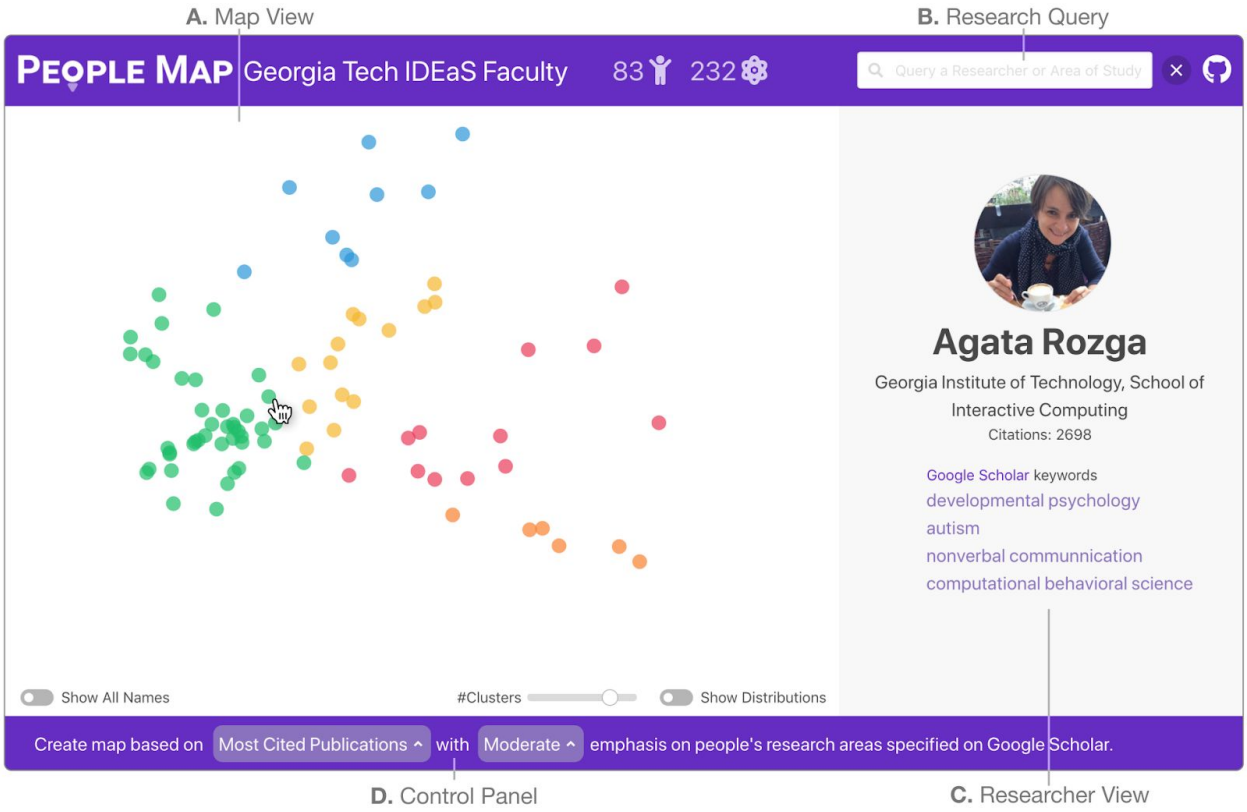

Figure 1: *Image of PeopleMap platform with labeled portions*

PeopleMap is an open-source, web-browser-based visualization tool that maps out researchers using natural language processing (NLP) techniques, allowing users to explore all the different information extracted from researchers' profiles using textual embeddings. It can determine the possible groupings of similarly-interested researchers, represent how researchers align with specified fields of study, and reveal potential Gaussian distributions describing the research topics present in the dataset.

PeopleMap's user interface consists of four major components:

1. **Map View** (Figure 1A) visualizes the research topic similarities among researchers

2. **Research Query** (Figure 1B) allows users to search for researchers and query areas of study
3. **Researcher View** (Figure 1C), which shows the detailed information of the researcher hovered over by the user (e.g., affiliation, citations, interests);
4. **Control Panel** (Figure 1D) allows users to adjust the hyperparameters of the Map View visualization.

Next, the following sections will describe each component in more detail.

Mapping Out Researcher Interests

The Map View (Figure 1A) of PeopleMap is a visualization of embeddings representing the researchers in the selected dataset. Within the Map View, each dot represents a researcher and their corresponding embedding projected into a two-dimensional space. With the researcher data extracted from Google Scholar, these embeddings were created using term frequency–inverse document frequency (TFIDF) matrices and principal component analysis (PCA), which is discussed in greater detail in the following sections:

Collecting Google Scholar data for each researcher

Generating a PeopleMap visualization requires only public data that anyone can access. We collect each researcher's public information from Google Scholar, which includes the researcher's profile, publications, and research interests using a Python-based module called **scholarly** (https://github.com/scholarly-python-package/scholarly). The specific information included are:

- Google Scholar profile URL

- Top 50 most cited publications (titles, abstracts, and years of publication)
- Top 50 most recent publications (titles, abstracts, and years of publication)
- Google Scholar profile keywords
- Citation count
- Institution affiliation
- Google Scholar profile photo

PeopleMap formats and stores all researcher data in a CSV file, one column for each category of information listed above.

Researcher Embeddings

Using the publication data extracted from Google Scholar, the title and abstracts of each researcher's publications are first concatenated together to create a combined document for each researcher. Additionally, Google Scholar keywords of each researcher can also be concatenated into their respective combined documents. After their creation, in order to normalize and prepare them for analysis, these combined documents are:

1. Removed words with non-English alphabet characters to restrict the bounds of analysis
2. Eliminated words with fewer than two characters in length to mitigate noise in the data
3. Converted words to lowercase to simplify capitalization
4. Cleaned of HTML tags
5. Cleaned of stop-words
6. Stemmed words to simplify syntax

Once the documents have been normalized, they are then converted into researcher embeddings representing each individual researcher through the use of the TFIDF technique.

This technique takes into account both the occurrence of each word within a researcher's publications and its frequency. Furthermore, it provides us a quantitative method by which we can ignore common words shared by most, if not all, of the researchers, while measuring specific important or characteristic words that differentiate researchers. Each researcher's embedding is a column in a TFIDF matrix, with each row representing the respective term values for a specific word in each researcher's embedding. The following equation represents the combination of **n** total researcher embeddings, each individually represented as vectors **v**, to create the combined TFIDF matrix **R**.

$$[v_1, v_2, ..., v_n] = \mathbf{R}$$

With the researcher embeddings in the TFIDF matrix, it is necessary to first reduce the dimensionality of the embeddings, which are vectors in a several-thousand dimensional space, so that they can be visualized. To achieve this, principal component analysis (PCA) is used to assist in feature extraction and elimination, simplifying the researcher embeddings into vectors within a two-dimensional space that can be visualized in the Map View (Figure 1A).

PCA was chosen as a starting embedding technique, because PeopleMap is one of the first tools for interactively mapping out researchers. Our primary goal is to create a platform that improves the explorability and interpretability of researcher datasets. While there are many potential embedding techniques for the textual data of researchers, the goal was to start with more classic embeddings that could provide adjustable parameters for the platform.

We purposefully used PCA over other potential visualization techniques, such as UMAP or t-SNE, because they tend to find structure within the noise of a dataset with small sample sizes compared to the dimensionality of the data, while PCA is well justified as a linear model for such

datasets. Thus, we use PCA since it fits the constraints of our researcher dataset and allows us to still find emergent patterns among the researcher embeddings. In the future, we endeavor to improve the complexity of our embeddings by exploring several potential embedding techniques.

Querying Researchers and Areas of Study

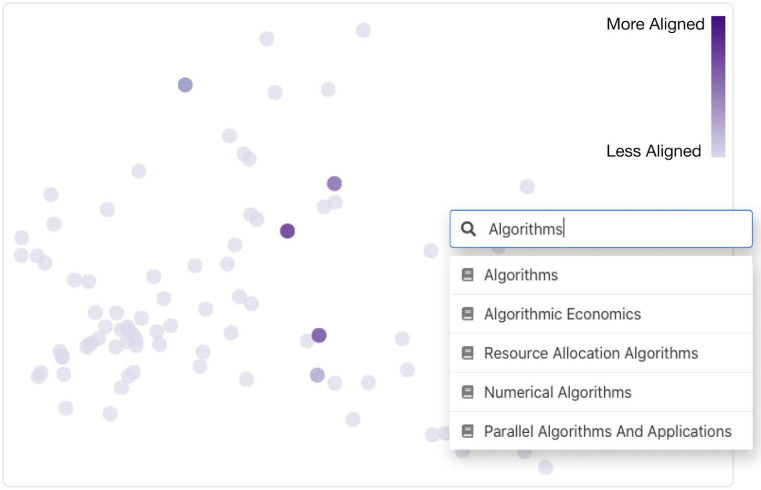

Figure 2: Research Query component and query results displayed in Map View. *Researchers are colored based on how well they align with the query (in this example, "Algorithms" is the query research topic)*

The Research Query component allows the user to both locate specific researchers, as well as see which researchers are aligned with each of the Google Scholar keywords collected from the researcher dataset. When the user searches for a researcher, PeopleMap highlights the researcher's representation in Map View by enlarging the dot's radius and outlining it; PeopleMap also displays the researcher's Google Scholar profile information in the Researcher View. When calculating a researcher's alignment with a selected Google Scholar keyword, PeopleMap uses similarity analysis between researcher embeddings and topic embeddings, which is discussed in the next section.

Similarity Analysis

The TFIDF researcher embeddings used for the Map View component are also used to calculate the similarity between a researcher and a specified topic. For example, if a user wants to see which researchers frequently use a specific term prominently throughout their work, it is possible to use their researcher embeddings to find which ones use the term most often compared to their overall writing. To calculate this, the specified topic (e.g. “natural language processing”) is first converted into a TFIDF embedding using the same process that is outlined for the researcher's publications.

Then, the cosine similarity between the specified topic embedding and each of the researcher embeddings in the TFIDF matrix is calculated, which indicates the similarity between the two vectors: the higher the value, the greater the similarity. The following equation represents the cosine similarity between the specified topic embedding, represented as the vector **q**, and the current researcher embedding, represented as the vector **v**, to produce the resulting similarity score, represented as **s**.

$$\frac{q \cdot v}{\|q\| \times \|v\|} = s$$

By performing cosine similarity calculations between the specified topic embedding and the researcher embeddings in the TFIDF matrix, the top similarity scores can be used to find the researchers that most align with the specified topic. These researchers are, in turn, highlighted in the Map View when the specified topic is queried in the Research Query (Figure 1B) component. Researchers are colored based on how well they align with the query. Darker indicates more aligned. The Research Query tool, together with the color gradient visualizing the query results,

help users better understand the scope of research relevance among the researchers. The researchers more prominently highlighted are those who tend to use the query term proportionally more than their peers in the dataset. This can serve as a reference to begin inquiries into the individual's research rather than serve as a full assessment of their contributions to that research topic.

Clustering Researchers

To help users more easily identify groups of related researchers, the Map View (Figure 1A) colors the researcher dots to indicate clusters of associated researchers. The intention of this coloration is not to create strictly-defined groups of researchers. Rather, we want to help users visualize the scope of shared interests. To assign these colorings, Gaussian mixture modeling was implemented, which will be explained in greater detail in the following section.

Assigning Colors

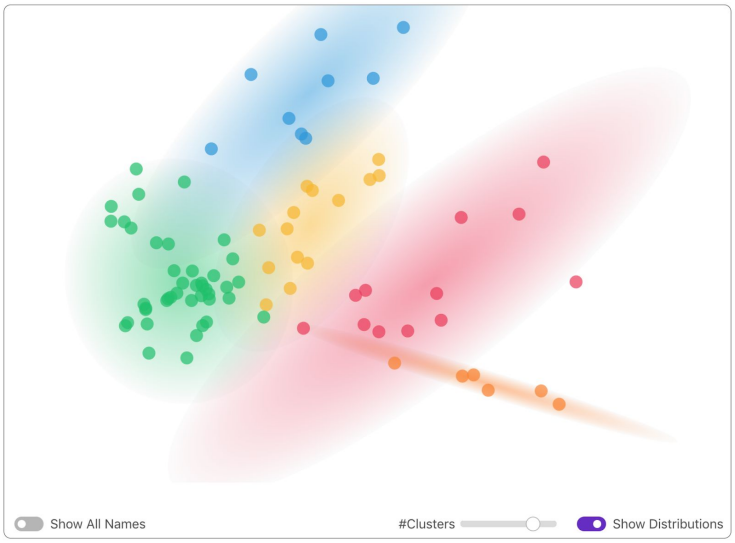

Figure 3: Map View component with cluster distributions displayed. Each dot is an embedding of a researcher's combined publications; the color is assigned by their membership in a cluster of a Gaussian mixture model, which are displayed as elliptical distributions. Proximity of dots indicates similarity of research interests while remoteness indicates disparity.

Previously, we used PCA to reduce the dimensionality of the researcher embeddings, projecting them into a two-dimensional space for visualization. This dimension reduction of the researcher embeddings is also necessary for clustering techniques to be performed. In the researcher dataset for the IDEaS faculty at Georgia Tech, the researcher embeddings have over 11,000 dimensions, with each dimension representing a word in the vast vocabulary shared by the researcher dataset; however, there are only 83 data-points. Thus, considering the complexity of the data, it is necessary to simplify the dimensionality of the data before performing clustering.

Therefore, using the newly-reduced researcher vectors created using PCA, the total set of researcher vectors is analyzed using Gaussian mixture modeling. Using this technique, the overall distribution of researcher vectors is categorized into several different Gaussian distributions. These distributions are meant to assist the user in their understanding of the different topics within the researcher dataset and how these topics are shared among different groups.

Once these researcher vectors are clustered using Gaussian mixture modeling, they are visualized within the Map View component of PeopleMap and colored according to their designated Gaussian distribution, with each distribution being assigned a unique color. Researcher dots that are close together tend to reflect a similarity in research pursuits between the two researchers; increased distance between researcher dots reflects the opposite. Using distance and coloring of a research embedding as metrics for gauging similarity, the user can

better understand the relationship between each of the researchers as well as the diversity of topics in the Map View.

Results/Discussion

The source code for PeopleMap is available at https://github.com/poloclub/people-map; it is registered under the permissive MIT license, making it available to anyone. It includes the PeopleMap visualization as well as the data collection and processing code for developing a new researcher dataset which can be loaded into the platform. Furthermore, our documentation provides concrete tutorial steps for users to follow, so that new users with beginner's experience in Python and Javascript may also easily set up the tool. It walks a new user through the initial steps of collecting data from Google Scholar to the final stages of setting up the PeopleMap platform on their computer.

In addition to the source code, we provide two live demos of PeopleMap that allow anybody to explore and become familiar with the PeopleMap platform. The first demo analyzes the publications of the faculty in Georgia Tech's Center of Machine Learning, while the second demo analyzes the publications of the faculty at the Institute for Data Engineering and Science (IDEaS), also at Georgia Tech. The corresponding datasets for these two faculty groups are available alongside the source code of the Github page.

Example Usage Scenario

James is an academic director at a university, looking to develop a new project centered around the study of black holes. He is looking for potential colleagues at his university with whom he can begin working on this new project. While he does have some current connections

with professors at his university, he would like to explore the diversity of researchers at his university by using PeopleMap.

To start, James clones the PeopleMap repository and begins following the steps of the documentation. Next, he goes to the university directory and gathers the Google Scholar profile names of all of the relevant researchers. Using tools included in the repository, he gathers their relevant publication information, processes the text, and generates the data files for the PeopleMap platform.

With PeopleMap fully set up, James begins exploring the researcher dataset with all the tools explained in the methodology. First, he uses the *Publication Set* drop-down in the *Control Panel* (Figure 1D) and selects *Most Recent Publications* since he wants to find researchers currently focusing on studying black holes. Next, James clicks the *Research Query* (Figure 1B) component and types "black holes", searching to see the researchers most closely aligned with the topic. The tool then highlights the top-five researchers associated with the topic. From this initial search, he discovers several individuals he did not know from his previous correspondence and decides to look a little deeper.

Using this information, James proceeds to use the *Researcher View* (Figure 1C) component to identify the researchers, clicking on their Google Scholar profile links to see some of their published work. However, before ending his search, he would like to see some of the other researchers that are in close proximity to the ones already selected. Using the *Keywords Emphasis* drop-down, he tries different choices of keywords to see the groups of researchers that emerge near the previously identified researchers, using the *Show All Names* toggle to take note of other researchers that are frequently associated with the ones found using the *Research Query*

component. With this wide array of researchers, James is confident he has gathered all the potential collaborators and proceeds to use their Google Scholar profiles found in the *Researcher View* component, as well as other resources, to gauge which ones would be the best fit for the project.

Scaling the Impact of PeopleMap

PeopleMap for research entities: PeopleMap could transform how research talents at research institutions may be summarized and discovered by both internal and external collaborators.

At Georgia Tech, we have successfully developed PeopleMap for two major research entities: IDEaS and the Center for Machine Learning. The leadership of IDEaS are very excited about this tool, especially the interactivity and explorability that it provides for researcher datasets as well as the ease with which it can be updated for new members. While we used the tool for faculty datasets in IDEaS and the Center of Machine Learning, it could be applied to the entirety of the College of Computing or even Georgia Tech as a whole. The scope of the researchers included is a matter of preference for the group seeking to implement PeopleMap.

PeopleMap for larger entities: Using the data-collecting and processing tools that are part of the PeopleMap repository, it is possible to expand the platform to other researcher datasets, as long as these researchers have Google Scholar profiles with their associated publications listed.

The PeopleMap for IDEaS visualizes 83 researchers. However, it is possible to have significantly more researchers than this amount; the limiting factor for the total count is essentially the size of

the Map View visualization. As more researchers are added, the higher number of dots can lead to greater visual complexity in the visualization, potentially causing "overplotting" as it becomes harder to distinguish between each of the dots and locate specific individuals using either the *Show All Names* toggle or the Researcher View component. Additionally, the researcher information within PeopleMap does not update automatically when researchers' Google Scholar profiles update. PeopleMap users would need to re-run the data collection and processing step to refresh PeopleMap.

PeopleMap as a complementary resource: Rather than replacing current directories, we developed PeopleMap as a tool to complement these existing directories. PeopleMap can be used in conjunction with the directories of universities, companies, agencies, and other institutions to lend an additional perspective upon the diversity of research interests that the institution holds.

Conclusion

PeopleMap, in its current form, will continue to be useful for years to come, but we plan on continuing to improve the system by increasing the sophistication of the NLP techniques used in analysis and expanding the available functionalities for exploring researcher datasets. In the current version of PeopleMap, we use TFIDF to generate researcher embeddings (discussed in Methodology section) from our gathered researcher data before using PCA and Gaussian mixture modeling for visualizing these embeddings and performing clustering techniques. However, as we seek to increase the complexity of our embeddings, we plan on exploring several potential embedding techniques. For example, we aim to extract hidden layers from pretrained and fine-tuned Transformer (Vaswani 2017) models such as BERT (Devlin 2018). Prior work has

explored fine-tuning these models on text data from the scientific domain, yielding improved results on downstream tasks (Beltagy 2019). However, we aim to use similar techniques in the context of visualization. Using these techniques, we open up the possibility of both improved information extraction and visualization of researcher datasets.

Lastly, we hope that PeopleMap can assist any individual seeking to delve deeper into the fields of interests found within any group of researchers. We encourage any institution composed of published researchers to use PeopleMap if they would like to explore the diversity of content produced by their members. We expect that recommendation systems for research papers and publication venues will continue to be a topic of interest in coming years, as there have been several different studies addressing potential platforms and solutions (Beel 2016, Medvet 2014, Beel 2017, Alhoori 2017, Küçüktunç 2013). Furthermore, we also expect organizations will seek to improve outdated directory systems so that both internal and external groups can more efficiently and confidently connect with researchers for potential collaborations.

References

1. Aizawa, Akiko. "An information-theoretic perspective of tf–idf measures." *Information Processing & Management* 39.1 (2003): 45-65.
2. Alhoori, Hamed, and Richard Furuta. "Recommendation of scholarly venues based on dynamic user interests." *Journal of Informetrics* 11.2 (2017): 553-563.
3. Beel, Joeran, et al. "paper recommender systems: a literature survey." *International Journal on Digital Libraries* 17.4 (2016): 305-338.
4. Beel, Joeran. "Towards effective research-paper recommender systems and user modeling based on mind maps." *arXiv preprint arXiv:1703.09109* (2017).
5. Belkhodja, Omar, and Réjean Landry. "The Triple-Helix collaboration: Why do researchers collaborate with industry and the government? What are the factors that influence the perceived barriers?." *Scientometrics* 70.2 (2007): 301-332.
6. Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text." *arXiv preprint arXiv:1903.10676* (2019).
7. Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." *Transactions of the Association for Computational Linguistics* 6 (2018): 587-604.
8. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
9. Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman. "Outta control: Laws of semantic change and inherent biases in word representation models." *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017.
10. Jones, Karen Sparck. "A statistical interpretation of term specificity and its application in retrieval." *Journal of documentation* (1972).
11. Kim, Han Kyul, Hyunjoong Kim, and Sungzoon Cho. "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation." *Neurocomputing* 266 (2017): 336-352.

12. Klein, Dan, and Christopher D. Manning. "Conditional structure versus conditional estimation in NLP models." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.
13. Krauthammer, Michael, and George Hripcsak. "A knowledge model for the interpretation and visualization of NLP-parsed discharged summaries." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001.
14. Küçüktunç, Onur, et al. "TheAdvisor: a webservice for academic recommendation." *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. 2013.
15. Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems*. 2001.
16. Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." *Pattern recognition* 36.2 (2003): 451-461.
17. Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.
18. McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).
19. Medvet, Eric, Alberto Bartoli, and Giulio Piccinin. "Publication venue recommendation based on paper abstract." *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. IEEE, 2014.

20. Ramos, Juan. "Using tf-idf to determine word relevance in document queries." *Proceedings of the first instructional conference on machine learning*. Vol. 242. 2003.
21. Sanz, Jose Antonio, et al. "IVTURS: A linguistic fuzzy rule-based classification system based on a new interval-valued fuzzy reasoning method with tuning and rule selection." *IEEE Transactions on Fuzzy Systems* 21.3 (2013): 399-411.
22. Tannier, Xavier. "NLP-driven data journalism: Time-aware mining and visualization of international alliances." *" Natural Language meets Journalism", workshop of the International Joint Conference on Artificial Intelligence (IJCAI 2016)*. 2016.
23. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
24. Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2.1-3 (1987): 37-52.
25. Xing, Chao, et al. "Document classification with distributions of word vectors." *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014.